DOCUMENT RESUME

ED 278 711 TM 870 141

AUTHOR	Boldt, R. F.; And Others
TITLE	The Validity of Various Methods of Treating Multiple SAT Scores.
INSTITUTION	College Entrance Examination Board, New York, N.Y.
REPORT NO	CBR-86-4; ETS-RR-86-25
PUB DATE	86
NOTE	13p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101 ($6.00).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Analysis of Variance; *College Entrance Examinations; College Students; Evaluation Methods; *Grade Prediction; Higher Education; Literature Reviews; Mathematics Tests; *Predictive Validity; Racial Differences; *Scores; Socioeconomic Status; *Test Theory; Verbal Tests; Weighted Scores
IDENTIFIERS	College Board Validity Study Service; Cross Validation; *Retesting; *Scholastic Aptitude Test; Student Descriptive Questionnaire

ABSTRACT

A review of the literature concerned with validity data and policies for various methods of treating multiple scores is reported, as are analyses of data from the College Board Validity Study Service. The analyses evaluated the use of Scholastic Aptitude Test--Verbal alone, Scholastic Aptitude Test--Mathematical alone, and the use of both in combination. The methods for treating multiple scores that were studied were to use the score from the administration with the highest Verbal plus Mathematical, the highest score, the most recent score, and the average. Best results, in terms of the highest average validity, were achieved using average Verbal plus Mathematical. Also evaluated were weighted combinations where all scores were used, but the highest or most recent received a unique, empirically determined weight. The combinations of variables and the four methods of treating multiple scores were evaluated in regression equations developed using one-time testers. All treatments of multiple scores resulted in underprediction of actual grades, with the highest score providing the least amount of underprediction. The discrepancy between predicted and actual grades varied across institutions. Data from the Student Descriptive Questionnaire were cross tabulated with the number of retests. It was found that first test score-level, income, and race were all related to the frequency of retesting. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED278111

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

P. K. Hendel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

X Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

College Board Report

No. 86-4

The Validity of Various Methods of Treating Multiple SAT® Scores

R. F. Boldt
J. A. Centra
R. G. Courtney

The Validity of Various Methods of Treating Multiple SAT® Scores

R. F. Boldt
J. A. Centra
R. G. Courtney

College Board Report No. 86-4

ETS RR No. 86-25

College Entrance Examination Board, New York, 1986

R. E. Boldt is Senior Research Scientist at Educational Testing Service.

J. A. Centra is Professor and Chair of Higher Postsecondary Education at Syracuse University.

R. G. Courtney is Senior Research Assistant at Educational Testing Service.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101. The price is $6.

Copyright © 1986 by College Entrance Examination Board. All rights reserved.
College Board, SAT, Scholastic Aptitude Test, and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Foreword 1

Abstract 1

Introduction 1

Related Program Research 2

Analysis of Validity Study Data 3

Discussion 7

References 8

Tables

1. Characteristics of Multiple Testers 4

2. Mean Validity Coefficients for the Score from the Largest V + M Total and the Highest, Latest, and Average SAT Scores 5

3. Variance Component Analysis of Validities 5

4. Validity Coefficients Using Composite Weights for Highest and the Sum of the Others, and Latest and the Sum of the Others, Compared with the Methods of Table 1 Using Cycle-81 Data 6

5. Variance Component Analysis of Average Discrepancies 6

FOREWORD

The search for unequivocal empirical evidence as to the best way to treat multiple sets of test scores of individuals in college admissions has spanned nearly a quarter of a century. This report summarizes the results of earlier studies and describes the results of a current analysis of data for 51,000 students at 87 colleges. The results of the present study were generally consistent with those of earlier studies that no method of treating multiple scores produced substantially higher validities than other methods. An exception to this trend should be noted in the United States Naval Academy study reported in 1982 and 1984.

The present study found that the highest correlation between SAT verbal and mathematical scores and college grade-point average was obtained by averaging the multiple sets of scores for students tested twice and three times. However, this finding was confounded by another that all treatments of multiple scores resulted in a slight underprediction of actual grades. The highest score provided the least amount of underprediction of actual college grades. Thus, when errors in prediction occurred, the errors tended toward a lower predicted college grade than that actually achieved.

Several methods of averaging multiple scores were also examined. The simple arithmetical average of all scores proved to be as valid as weighted averages. The authors observe that, "From the point of view of validity, therefore, it would seem that the best choice is the average because it is the simplest to use, because it is as valid as the more complex weighting schemes, and because it is more valid than using the highest and latest scores."

What guidance can this study give for the treatment of multiple test scores? On the basis of predictive validity, the arithmetical average proved to have the highest correlation with college grade-point average. Thus an average of scores can be recommended for use for reasons of both validity and simplicity.

All treatments of multiple scores – average, latest and highest – slightly underpredict college grade averages. The discrepancy between predicted and actual grades varied sufficiently across colleges to suggest strongly that each institution should perform validity studies in order to determine whether the discrepancy is large enough to take into account in admitting students with multiple scores.

The results of this study were obtained from data supplied by 87 colleges. Whether similar results will pertain to any one college can only be determined by a validity study for that college. The College Board invites colleges to use its Validity Study Service for this purpose.

In addition to research outcomes, colleges may wish to be guided by considerations of equity in the admissions process. Table 5 of this report shows that ethnic minority and lower income students are less likely to repeat the tests. If colleges use the highest scores, they may be advantaging those who can afford to repeat the tests to the disadvantage of those who cannot.

Robert G. Cameron
Executive Director
Research and Development

ABSTRACT

A review of the literature concerned with validity data and policies for various methods of treating multiple scores is reported, as are analyses of data from the College Board Validity Study Service. The analyses evaluated the use of SAT-V alone, SAT-M alone, and the use of both in combination. The methods for treating multiple scores that were studied were to use the score from the administration with the highest V + M, the highest score, the most recent score, and the average. Best results, in terms of the highest average validity, were achieved using average V + M. Also evaluated were weighted combinations where all scores were used, but the highest or most recent received a unique, empirically determined weight. The use of the empirically determined weights was not superior in cross-validation.

The combinations of variables, and the four methods of treating multiple scores that did not involve the determination of empirical weights, were evaluated in regression equations developed using one-time testers. All treatments of multiple scores resulted in underprediction of actual grades, with the highest score providing the least amount of underprediction. However, the discrepancy between predicted and actual grades varied greatly across institutions.

Data from the Student Descriptive Questionnaire (SDQ) were cross tabulated with the number of retests. It was found that first test score-level, income, and race were all related to the frequency of retesting.

INTRODUCTION

Though College Board retesting and score-reporting policies have long been in existence, the fairness and effectiveness of these policies are periodically reconsidered. Three questions are of particular interest at the present time: What advice can be given to user institutions about the best use of multiple scores? What combination of predictors is most valid? What combination of SAT-V and SAT-M might be most useful in score reports? With each reconsideration it has become apparent that reaching and rationalizing any choice of policy or practice entails some appeal to empirical data. Hence there exist a number of studies bearing on the use of multiple scores. Some of these studies were conducted by the Board, but other test programs have also gathered relevant evidence as well. These studies, as well as analyses of the College

Board Validity Study Service file, are the resources used in the study to answer the questions above.

RELATED PROGRAM RESEARCH

In 1958 Olsen and Schrader reported on a study that compared the validities of various combinations of junior and senior scores for students who had completed two college years. The 3,368 students in the sample were selected from nine participating College Board institutions (three men's liberal arts, three women's liberal arts, and three engineering colleges). All the students had taken the SAT as seniors in 1954, with a subset of 1,650 that were initially tested as juniors in 1953. Students with more than two scores or with any testing pattern other than that described above, were omitted from the sample. The authors used an analysis of covariance to compare predictive values of a single test versus multiple test scores. In addition, comparative validities of various combinations of junior and senior scores were reported that included the junior and senior scores alone, the higher score, the weighted average (in which the senior score is counted twice), and the best-weighted combinations determined separately for each college. The students in the study had completed two years of college work and the analysis was based on the two-year cumulative grade-point average. The validities of junior and senior scores for each of the nine colleges were computed.

The authors confess to a consistent absence of significant results from the comparison of validities of the different combinations. "Altogether the results ... indicate that for candidates tested twice, preliminary and final scores differ little in validity." "Validity of higher score is not much better — or worse — than validity of either of the separate scores." Olsen and Schrader also reported no clear advantage for any of the combination methods used. "Combining the two scores with optimal weights resulted in a slight increase in validity, but the advantage is probably too small to warrant combining the preliminary and final scores."

Duggan (1959) used the Olsen and Schrader study to support the claim that "junior scores have virtually the same validity in predicting college grades as the senior scores." He recommended that colleges use the weighted average over any single score method and provided formulas for obtaining this combined score, which was simply the average of twice the senior score plus the junior score. For students studying mathematics as seniors, adding 10 additional points was suggested.

Two research reports (Olsen and Schrader 1959a and 1959b) described a continuation of the Olsen and Schrader work that used the same sample of colleges but with four-year grade-point averages. Junior and senior SAT scores were compared by counting the number of colleges for which each score yielded a higher validity. Neither score was proven superior in this four-year study. Citing Duggan (1959), and noting a College Board preference for the weighted average, Olsen and Schrader recommended use of this method. Their results support this conclusion only very slightly.

Watkins and Schrader (1963) examined the usefulness of the difference between junior and senior scores for predicting college grades with the sample used in the Olsen and Schrader studies. The difference between the two scores was found to be generally not significant and use of the difference for predicting college success was not recommended.

Cowen and Abrahams (1982) compared the validity of the average of multiple admissions test scores (AVECAT) with the highest score for 588 U.S. Naval Academy students. Both SAT and ACT scores were used, but ACT scores were converted to SAT equivalencies. A more detailed report of this study appears in an unpublished paper by McNitt (1984), who also participated in this research. McNitt explained that "separate analyses were made initially for SAT-verbal and SAT-mathematical and ACT English and mathematics scores, but the results from both analyses were similar, permitting the ACT scores to be converted to the SAT 200–800 range using the USNA equivalency table." Applicants to the Naval Academy are permitted to take as many entrance exams as they wish, but all test scores must be reported. McNitt explained that in a recent year, the number of multiple test scores for an examinee ranged from 2 to 19 with 18 percent submitting four or more sets of test scores. Cowen and Abrahams did not indicate how many tests were taken by students in the sample. Students were ranked within the sample on both highest and average scores. The average was found to be a better predictor of first-year grade-point average and was recommended as a more appropriate method for treating multiple scores. The correlation between average scores and first-year grades was .48; for highest scores and grades it was .38. "A better qualified group of applicants would be selected for admission to the Naval Academy if AVECATS were used as the selection score."

According to McNitt, the administration of the Naval Academy has already begun to use the average of multiple scores in the admissions equation. The new policy would use a straight average of the submitted scores, but would also allow a very low score to be dropped.

Pitcher (1965) computed validity coefficients for 111 LSAT repeaters, who were part of a study of almost 2,200 law school entrants. The sample consisted of 69 percent of all LSAT repeaters on the rosters of the 10 law schools participating in the study. The repeaters were less than 15 percent of the examinees; all but seven of these repeated the test once. Pitcher found that the average of the LSAT scores was not consistently better as a predictor than the earliest score alone, although either of these was consistently better than the latest score. Pitcher recommended that law schools use the average LSAT for prediction.

Pitcher (1977) also reported results from different treatments of multiple test scores for two separate studies. The first study conducted in 1974 examined the validity coefficients for a sample of 17,802 law students in 25 schools. This sample included 2,510 test repeaters — 14

percent of the total. The second study (1976) looked at regression equations and the prediction of law school grades. Unlike the first study, which included all law schools with validity studies, the second study considered only those schools with 50 or more repeaters. Comparisons were made across several combinations of scores (e.g., average, latest, initial). The results were not definitive. Preference was given to the average because it is simple to use and easier to explain.

Boldt (1977) evaluated the predictive accuracy of three methods for treating multiple LSAT scores. Several figures of merit (e.g., validity, error of prediction, standard deviation of the residuals, mean residuals) were used to evaluate the effectiveness of the treatments. Analysis of variance of the figures of merit revealed that the effects attributable to the methods of combining scores were quite small. Because the various methods were about equally effective, Boldt concluded that the average should be recommended for treating multiple scores since it is easy to use and widely understood.

In a position paper that considered the issues related to use of multiple test scores, Linn (1977) reviewed the literature on multiple LSAT scores. His review focused on the studies conducted by Pitcher (1977) and Boldt (1977). Linn explained that despite the facility of the average, some problems existed in that the average tended to underpredict. This effect was not large, and proved to be inconsistent from school to school (i.e., in 7 of the 25 schools repeaters were overpredicted). He concurred with the recommendations of Pitcher and Boldt: the results of these studies do not justify implementing a more elaborate procedure since the straight average is easier to use and more readily understood.

Multiple test scores present less of a problem at the graduate level for at least two reasons. First, there are fewer students who repeat (Powers, Clark, and Grandy [1984] reported that only 11 percent of the GRE test takers state they are repeaters), and second, the population taking the graduate examinations is both more selective and more stable than SAT takers. In the LSAT studies described above, there is agreement that the best method of treating multiple scores is the straight average, unless there is reason to believe that test performance is affected in some way by illness, or some other negative factor.

The PSAT/NMSQT and LSAT studies were limited in the number of institutions and students studied, and therefore in their conclusions. Olsen and Schrader studied only nine colleges, for which 49 percent of the 3,368 examinees had taken the SAT in their junior and senior years. Pitcher and Boldt studied no more than 25 law schools, at which less than 15 percent of the students were LSAT repeaters. The Naval Academy results, which showed the greatest discrepancy between highest and average score validities, were based only on a single sample. The purpose of the present study was to investigate further the various methods of treating multiple SAT scores with a more extensive sample of colleges and students.

A survey conducted by Novaks (personal communication 1984) found that colleges used a variety of practices with multiple test scores. The highest score (e.g., the highest V from any administration), the most recent score (e.g., the most recent V), the score from the administration with the highest V + M total (e.g., the V score from that administration), and the average score were among those used (e.g., the average of all the examinees' verbal scores). The preference emerging in the literature reviewed above is for the average or a weighted average.

ANALYSIS OF VALIDITY STUDY DATA

The colleges studied were drawn at random from those participating in the College Board Validity Study Service. The studies were conducted using students entering college in 1980 (cycle 80) and 1981 (cycle 81). The sample was restricted to those colleges with at least 100 cases with complete data consisting of SAT-V, SAT-M, high school average or high school rank, and responses to the Student Descriptive Questionnaire. Equal numbers of schools were selected from cycle 80 and cycle 81. (For half the schools in each cycle the high school performance variable was a grade, for the other half it was a rank in class. This variable played a role in the study for which the sample was drawn [Boldt 1986], but is not used in the present study.) The cases in this sample were matched to the Admissions Testing Program history file in order to obtain the complete testing history of the examinees. Thus, multiple test data could be brought together with the first-year averages available from the Validity Study Service. Additional schools were eliminated from the sample if they had fewer than 50 students who had taken the test twice. The resulting sample consisted of data from 87 colleges. Validity data and data from the Student Descriptive Questionnaire were available for analysis. The validity analyses were conducted to find the most effective method of treating multiple scores. The analysis of the SDQ data was accomplished to determine whether, as is commonly supposed, parental income level, performance on a first administration of SAT, and ethnic group are related to retesting behavior.

Analyses of Student Descriptive Data

What can be said about the demographic and academic characteristics of multiple testers? Is there a relationship between those who take the SAT more than once and parental income, student score level, high school grades, sex, or ethnic background? These questions were investigated using cross-tabulations of numbers of times tested with responses to the Student Descriptive Questionnaire, which is completed by most examinees as part of the registration procedure. Results of selected two-way tabulations are presented in Table 1.

Major differences can be noted for all categories except sex, for which similar proportions, 59 percent of the males

Table 1. Characteristics of Multiple Testers

		Number of Times Tested		
	N	*1*	*2*	*3*
Score Level				
(Based on Earliest V + M)				
1,200	4,488	55*	41	4
1,000 to 1,200	15,041	40	52	8
850 to 1,200	15,928	36	54	10
850 or below	15,514	39	49	12
Parental Income				
50,000 or more	5,477	36	52	12
24,000 to 49,900	16,715	39	52	9
9,000 to 23,900	15,438	43	50	7
Under 9,000	3,045	52	43	5
Self-Reported Grades				
3.5 to 4.5	18,301	41	50	9
2.5 to 3.5	25,244	38	53	9
0 to 2.5	3,376	46	46	8
Sex				
Males	23,615	41	49	10
Females	27,356	39	52	9
Ethnic Background				
American Indian	192	52	41	7
Black	3,254	54	41	5
Mexican-American	603	64	33	3
Asian-American	1,712	42	47	11
Puerto Rican	292	48	44	8
White	39,467	38	53	9

*Entries under *Number of Times Tested* are percentages, and add horizontally to unity.

and 61 percent of the females, were multiple testers. By score level, for which the earliest V + M was used to classify students, those at the highest score level were least likely to take the SAT more than once: 45 percent of the students at or above the V + M level of 1200 took the test either two or three times, compared to 60 percent of those at the 1000 to 1200 level, 64 percent of those at the 850 to 1000 level, and 61 percent at the 850 and below level.

Parental income was also related to the number of times tested: 64 percent of those in the highest income level – over $50,000 – tested more than once, compared to 48 percent of those in the under $9,000 parental income category. Multiple testers also increased at each income level, with 57 percent of those in the $9,000 to $23,000 category and 60 percent of those in the $24,000 to $49,000 category testing more than once.

Because students from higher income families, as a group, receive slightly higher test scores (Donlon 1984), additional cross-tabulations were run to investigate further the interaction of score level and income on test-taking frequency. The ethnic background of the student was also included in a four-way tabulation. Students at the highest score level (V + M greater than 1200) have similar test-taking patterns regardless of income or race. For example, just over 40 percent of white students with scores over 1200, at all income levels, took the SAT either two or three times. But at lower score levels, income was definitely a factor. In the 1000 to 1200 score range for white students, 48 percent of those from higher incomes were multiple testers compared with 36 percent of those from lowest-income families. In the 850 to 1000 score range, 44 percent of high-income students and 31 percent of low-income students were multiple testers. Similar patterns, though with different percentages, were observed for students from other ethnic groups. Thus for all racial groups and at all but the highest score level, students from higher-income families take the test more frequently.

As further shown in Table 1, students in the lowest range of self-reported high school grades tested less frequently than those with higher grades: 54 percent of the students in the 2.5 and below range tested two or three times, compared to 59 percent of the 3.5 to 4.5 students, and 62 percent of those in the 2.5 to 3.5 grade-point range. Thus, students with the poorest high school records – defined by the approximately 7 percent of this sample who reported grades less than 2.5 – are least likely to test more than once.

Finally, in Table 1, the test-taking frequencies of students in each of six ethnic groups are given. White students repeat the SAT more than students from other ethnic groups: 62 percent take the test either two or three times. Asian-American test-takers follow closely behind with 58 percent repeaters. Least likely to repeat were Mexican American students, with only 36 percent taking the test two or three times. Black, American Indian, and Puerto Rican students fell between the extremes, with 46, 48, and 52 percent repeaters, respectively.

The lower repeater rate among Mexican American students may be because of their geographic location. Most are in the Southwest and attend less competitive postsecondary institutions. The lower proportion of black student repeaters, according to the four-way tabulations, is especially noteworthy in the low-score/low-income categories. Middle-income black students who scored initially in the 850 to 1200 range repeated the test at rates similar to Asian-American students (56 percent) and just below the white rate (62 percent).

Predictive Validity Analyses

These analyses were intended to compare the average validity of the methods Novaks (personal communication 1984) had found in use. These were the highest score, the most recent, the score from the administration with the highest V + M total, and the average score; previous studies employed most of these methods. The results of the analysis are contained in Table 2, which presents mean validities for those tested two and three times. Less than 2 percent of SAT

Table 2. Mean Validity Coefficients for the Score from the Largest V + M Total and the Highest, Latest, and Average SAT Scores

	Score from Highest V + M Total	*Highest*	*Latest*	*Average*
Students Tested Twice (87 Schools; 25,770 Students)				
SAT-V	.33	.34	.33	.34
SAT-M	.32	.32	.31	.33
SAT-V + M	.38	.38	.37	.39
Students Tested Thrice (24 Schools; 3,561 Students)				
SAT-V	.20	.21	.20	.22
SAT-M	.22	.23	.22	.24
SAT-V + M	.26	.25	.25	.27

Table 3. Variance Component Analysis of Validities

Source	*Tested Twice (Percent)*	*Tested Thrice (Percent)*
Institution (I)	78.7	88.1
Variables (V)	5.2	1.4
Method (M)	.3	.1
I × V	14.5	8.1
I × M	.7	1.6
V × M	.0	.0
I × V × M	.6	.7
Total	100	100

repeaters were tested more than three times. The process for averaging validities involved the use of a form of Fisher's z that uses an improvement suggested in Kruskal and Tanur (1978) and is the process used in developing all other average validities given in this report. The values of the entries in the table are consistent with two principles: (1) V + M is better than V or M alone regardless of the treatment of multiple scores, and (2) averaging is slightly the best of the four treatments of multiple scores regardless of the combination of V and M used; the increase for the average over other treatments is .01 to .02.

Because the entries in Table 2 are averages of within-school analyses, there remains the possibility that there are institutions for which highly atypical patterns exist. The Naval Academy study, for example, indicated that average scores produced considerably higher validities (Cowen and Abrahams 1982). Therefore the interactions of institution with the method of treating the multiple scores was examined, as were other components of validity coefficient variance. Note that there are three combinations of variables, V, M, and V + M, and four treatments of multiple scores, latest, highest, average, and score from administration with highest V + M. This makes the 12 combinations of variable and method that correspond to the entries in Table 2. To construct Table 2 the 12 validities were developed for each college, for a total of 1,044 and 288 coefficients for the two- and three-time testers, respectively. Variance component analyses of these validities were conducted to determine the extent to which variation in the coefficients was associated with the interaction of institution with method, to see if the efficacy of the method of treating multiple scores varied greatly according to the institution*. The results of the analyses are presented in Table 3. Note that less than 2 percent of the total variance is associated with the interaction of institution by method (I × M). It would appear, therefore, that the conclusions from Table 2, i.e., that the validities are similar regardless of the method of treating multiple scores, can be accepted as applying to particular institutions. As a further check on this conclusion the actual table of interactions of institution with method was examined to see if perhaps some one or two institutions differed greatly from the rest. Such was not the case. Thus, the greatly superior validity of averaging the scores that was observed at the Naval Academy was not found at any other institution.

Note a substantial institution by variable interaction (I × V) in Table 3. The nature of this effect became clear when the computed interactions were examined: 99 percent of the interaction was due to variation in the balance of the validity of V (verbal) alone with that of M (mathematical) alone. In other words, for some institutions the validity of V was greater relative to that of M than at other institutions. The following computations help evaluate the magnitudes of these differences. When the effect of the interaction of variable with institution was added to the main effect of the variables used, the largest value was associated with the sum of V and M for 62 of the 87 institutions with data for two-time testers, and for 11 of the 24 institutions with data for three-time testers. For those institutions in which the validity of V was largest, its average advantage over V + M rounded to zero for both two- and three-time testers; for those institutions for which M was largest, its average advantage over V + M was .006 for the two-time testers, and rounded to zero

*The variance component analyses used in this study are not traditional analyses of variance, which apply distributional theory to test hypotheses. They merely describe the variation of the particular coefficients with which Table 3 and Table 5 are concerned. Take, for example, the validities for two-time testers in Table 3. One could draw a number of validities, one at a time and with replacement, and calculate a variance of the values drawn. It can be shown that the expectation of this variance is the unweighted sum of suitably defined variance components, and in that sense the components account for percentages of variance. The "suitable definitions" are as follows: (1) the coefficients can be represented as an additive model with the usual terms represented in a complete factorial experiment, (2) all components are fixed effects, i.e., there is no sampling of effects, and (3) the variance components are defined, for a given component, as the average of the squared parameters of the component. That is, the divisor in the mean square is not the degrees of freedom but the actual number of parameters. For example, in the I × M interaction, the variance component equals the sum of squares divided by the 348, that is the number of institutions, 87, times the number of methods, which is 4. This representation is purely descriptive, makes no assumptions about the distributions of scores or grades, and does not involve significance testing. Hence the formats of tables 3 and 5 involve only percents of variance accounted for.

for the three-time testers. Hence, even though V + M did not always afford the greatest validity, its validity was not exceeded by much, and on the average it did the best. Apparently, V + M is almost always the best choice, or very nearly the best choice.

The policies examined in tables 2 and 3 do not exhaust those mentioned in the literature review. Another policy was that mentioned by Duggan (1959), who recommended the average of twice the senior score plus the junior score. This approach can be generalized to the use of a particular weight for the most recently obtained score and a different weight for the others. Similarly, a weighting of the highest score and other score(s) could be used. Such policy will be referred to as an "unequal weight method." An overall policy of this type could take the form of a recommended ratio of the two weights for all institutions. The ratio would probably be unique to the number of tests taken. A procedure for empirically determining the optimum ratio of weights has been developed by Boldt (1975). In the present study this procedure was used to generate weights both for two- and three-time testers for the cycle-80 data; the resulting relative weights were then used in the cycle-81 data, thus eliminating capitalization on chance that would occur if the evaluation of the weights were done using the same sample on which they were developed. Table 4 presents the cycle-81 average validities that are analogous to those in Table 2 for both cycles, and, in parentheses, the validities obtained using the empirical unequal-weight method. Note in Table 4 that for any combination of variable and number of tests taken (line), the validities obtained using the empirical weights (values in parentheses) are very similar regardless of whether the unique weight is given to the highest or the most recent score. The largest such difference is between .24 and .22 in the last line of the table, which occurred using both V and M with three-time-tester data. Note also that, for any particular combination of variables and number of retests (line), the highest validities were obtained using the empirically weighted composite or the average, e.g., on the last line of the table the highest values are .24, .22, and .23. Note that there is no combination of number of retests and combination of variables (line) for which a validity of an empirically weighted composite (value in parentheses) excedes the validity of the simple average (value in the far-righthand column). In fact, the .24 in the last line of the table is the only one for which the validity of the simple average is not among the highest in the line. Thus the use of the unequal-weight policy did not lead to improvement over the easily applied and understood procedure of averaging scores, and was not pursued further in this project.

We have evaluated the various methods of using multiple scores in terms of their average validities. Another way to evaluate them is to see how their behavior in a predictive regression formula compares with that of the single test. One might expect, for example, that the highest score from a set might bear a different relationship to a criterion measure than would a single, first-time observation.

Table 4. Validity Coefficients Using Composite Weights* for Highest and the Sum of the Others, and Latest and the Sum of the Others, Compared with the Methods of Table 1 Using Cycle-81 Data

	Score from Highest V + M Total	*Highest*	*Latest*	*Average*
Students Tested Twice (44 Schools; 11,929 Students)				
SAT-V	.34	.34(.35)	.33(.35)	.35
SAT-M	.30	.30(.32)	.29(.32)	.32
SAT-V + M	.37	.37(.38)	.36(.38)	.38
Students Tested Thrice (13 Schools; 1,854 Students)				
SAT-V	.17	.16(.17)	.16(.17)	.17
SAT-M	.18	.18(.21)	.20(.20)	.21
SAT-V + M	.21	.20(.24)	.21(.22)	.23

*The weights were developed in cycle 80 and the validities of weighted composites are in parentheses.

Table 5. Variance Component Analysis of Average Discrepancies

Source	*Tested Twice (Percent)*	*Tested Thrice (Percent)*
Institution (I)	94.1	88.1
Variables (V)	.4	2.8
Method (M)	2.6	5.2
I × V	2.3	2.7
I × M	.4	.9
V × M	.1	.1
I × V × M	.1	.2
Total	100	100

To test this, sets of "discrepancies" were developed that could indicate systematic errors of prediction by the different methods of treating multiple scores using the following steps: (1) the formula for the regression of first-year average on test score was developed for the nonrepeating examinees (one formula per school); (2) the score obtained by a method being evaluated was substituted into the regression formula of step (1) to obtain a predicted first-year average for that examinee if the method were used; (3) the examinee's actual first-year average was subtracted from the predicted first-year average to obtain a discrepancy score for that examinee for that method, which is positive if the prediction was too high, and negative if the prediction was too low; and (4) the discrepancies were averaged for the institution. Because the criterion scale is a factor in these discrepancy data, colleges that did not grade on the four-point scale were dropped, leaving 83 colleges as sources of data for two-time testers, and 23 colleges for three-time testers. As with tables 2 and 4, there were 12 combinations of method and variable, resulting in 996 and 276 coefficients for the two- and three-time testers, respectively. Variance component analyses like those leading to Table 3 were conducted using the average discrepancies. The results are presented in Table 5. Clearly the major source

of variance in the tables of average discrepancies is the institution (INST). The sizes of the institution effect range from − .43 to .32 on a four-point scale for the two-time testers, and from − .53 to .21 for the three-time testers. The methods main effect (METH) is the second most potent source of variance, with the treatment effects being .01, .03, − .01, and − .03 for score from highest V + M, highest, most recent, and average methods respectively for the two-time testers. The corresponding figures for the three-time testers were .02, .04, − .01, and − .05. Note that the methods that involve choosing the highest of a set tend to overpredict relative to the most recent and average, as one would expect. Using the most recent score produces a higher prediction than the average, probably because receiving an initial unexpectedly low score is among the reasons for repeating the test. The interactions with method (I × M, V × M, and I × V × M) produce negligible variance.

One finding of interest is that the overall mean of the average discrepancies is negative, being − .04 for the two-time testers, and − .03 for the three-time testers. If these averages are added to the method's main effects, the resulting values are all negative, hence the least biased of the methods is to take the highest score.

DISCUSSION

In this study, five methods of treating multiple scores were investigated, of which the two averaging methods are discussed first. Some form of averaging should be a strong candidate for a method of treating several scores because it seems less prone to errors of measurement than do others that select extreme scores. In this study, three kinds of averaging were investigated: giving all the scores the same weight, giving all the scores except the latest the same weight, and giving all the scores except the highest the same weight. Using the highest and the latest scores can be regarded as arbitrary weighting schemes in which the scores other than the highest or latest are given weights of zero. In contrast, computed weights were chosen that provided least squares fits in the cycle-80 data, and then were validated in the cycle-81 data to avoid capitalization on chance. As one could expect, the variables calculated using the weights developed empirically proved to be more valid than the variables computed using the arbitrary zero weights, i.e., the highest or most recent. Therefore, if one were selecting treatments for multiple scores, one might favor the empirical weighting scheme. However, that scheme produced average validities no better than those achieved by the simple average. From the point of view of validity, therefore, it would seem that the best choice is the average because it is the simplest to use, because it is as valid as the more complex weighting schemes, and because is more valid than using the highest and latest scores.

From the validity point of view, then, using the average of all scores seems to be the best method for treating multiple scores. But from the point of view of effects on prediction, the average seems to be the worst. Its treatment effect is the smallest, and the overall average discrepancy was negative. If the overall average is added to the main effect of method of treating multiple scores, the sums are − .06, − .04, − .08, and − .11 for two-time testers, and − .03, − .01, − .07, and − .11 for the three-time testers. (It will be remembered that these figures are on a four-point scale.) The order of the treatments for these figures are the score from highest V + M, highest, latest, and average. Hence taking the highest score seems to be the best simply because all the methods tend to underpredict.

From the data in the present study, the decision as to which is the most preferable treatment of multiple scores seems to depend on how one evaluates the discrepancy differences as compared to the validity differences. The study itself does not supply a method of making that evaluation. But the study does establish that in choosing methods one is not risking, for example, the loss of 15 to 20 validity points or a bias of a quarter to a half a point on a four-point grade scale.

A very striking feature of the data is the variation in the size of institution (INST) main effects in the analysis of discrepancy scores. The range from smallest to largest is almost three-fourths of a grade point for both the two- and three-time testers. It is not clear what the reason is for this variation, or whether the effect would prove reliable in repeated studies with the same institutions.

It might be expected that there would be definite differences in the characteristics of multiple testers, and indeed we found them in this study. Higher-income students tend to test more frequently than lower-income students, and this is true not only for white students but for all ethnic groups. Indeed, family income tends to be related to test-taking frequency for all but the very highest scoring students; students above the 1200 V + M score range tend to repeat the SAT less than lower-scoring students regardless of their income or ethnic group.

Students who score high on the SAT at first testing who choose not to retake the test are apparently satisfied with their performance and believe that their scores will not improve significantly with repeated testing. For the most part they are correct, given regression effects and other factors (Donlon 1984). Lower-scoring students, however, are more likely to retake the test because, most likely, they believe their scores can improve. Parental income may relate to test frequency because of the likelihood that students from high-income families can apply to a wider range of colleges, since costs are somewhat less critical, and higher scores will improve their chances of acceptance at the more selective institutions.

The ethnic difference in test frequency may be related in part to geographic location, as in the case of Mexican-American test-takers, and in part related to parental income and other characteristics. Low-scoring black students from low-income families, for example, probably have more limited college plans than middle-income/middle-scoring

black students who, as with Asian-American students, retake the SAT almost as frequently as white students.

REFERENCES

Boldt, R. F. 1975. Comparison of a Bayesian and a Least Squares Method of Educational Prediction. Educational Testing Service Research Bulletin 75-5. Princeton, N.J.: Educational Testing Service.

Boldt, R. F. 1977. Evaluation of Three Methods for Treating Repeaters' Scores. Report LSAC-77-7. In *Law School Admission Council, Reports of LSAC Sponsored Research*, Vol. 3, 1975–77. Princeton, N.J.: Law School Admission Council.

Boldt, R. F. 1986. Generalization of SAT Validity Across Colleges. College Board Report 86-3 (ETS Research Report 86-24). New York: College Entrance Examination Board.

Cowen, M. B., and N. M. Abrahams. 1982. *Selecting Qualified Candidates to the United States Naval Academy Using College Aptitude Test Scores*. NPRDC Special Report 82-20. San Diego, Calif.: Navy Personnel Research and Development Center.

Donlon, T. R. 1984. *The College Board Technical Handbook for the Scholastic Aptitude Test and the Achievement Tests*. New York: College Entrance Examination Board.

Duggan, J. M. 1959. Puzzles and Powers in Junior SAT Scores. *College Board Review*, Winter pp. 37–40.

Kruskal, W. H., and J. M. Tanur eds. 1978. *International Encyclopedia of Statistics*. New York: The Free Press.

Linn, R. L. 1977. On the Treatment of Multiple Scores for Law School Admission Test Repeaters. Report LSAC-77-7. In *Law School Admission Council, Reports of LSAC Sponsored Research*, Vol. 3, 1975–77. Princeton, N.J.: Law School Admission Council.

McNitt, R. W. 1984. Highest SAT or Average SAT. U.S. Naval Academy, (unpublished manuscript).

Novak, T. A. 1984. Personal communication.

Olsen, M., and W. B. Schrader. 1958. *Relation of Preliminary and Final Scholastic Aptitude Test Scores to College Grades: Preliminary Report*. College Entrance Examination Board Research and Development Reports; also Statistical Report 58-15. New York: College Entrance Examination Board.

———. 1959a. *The Use of Preliminary and Final Scholastic Aptitude Test Scores in Predicting College Grades*. College Entrance Examination Board Research and Development Reports; also Statistical Report 59-19. Princeton, N.J: Educational Testing Service.

———. 1959b. *The Comparative Achievement of Final-only and Preliminary-final Scholastic Aptitude Test Candidates*. College Entrance Examination Board Research and Development Reports; also Statistical Report 59-20. Princeton, N.J.: Educational Testing Service.

Powers, D. E., M. J. Clark, and J. Grandy. 1985. Test Score Changes on the GRE General (Aptitude) Test. Educational Testing Service Research Bulletin 85-4. Princeton, N.J.: Educational Testing Service.

Watkins, R. W., and W. B. Schrader. 1963. The Relative Differences Between Junior and Senior Scholastic Aptitude Test Scores and College Grades. College Entrance Examination Board Research and Development Reports; also Statistical Report 63-29. Princeton, N.J.: Educational Testing Service.